

# The Efficacy of Collaborative Learning Recitation Sessions on Student Outcomes \*

Kim P. Huynh, David T. Jacho-Chávez and James K. Self

December 22, 2009

For most students, freshmen-level introductory economic courses represent their first exposure to the science of economics and its powerful tools of reasoning. These tools are the greatest benefits afforded to students who gain at least a principle level understanding of economics. However, here is the problem: for most universities and colleges, students in several majors will be required to take at least one course in economics and will generally take that course in their freshman year. Anecdotally, there is little enthusiasm for learning economics and many express fear that the course will be excessively difficult. One way to help overcome these obstacles is to incorporate opportunities for Collaborative Learning (CL) into the pedagogy. This addresses a criticism that undergraduate economic courses are too reliant on passive learning (lecture-only classroom activity), which was raised by [Becker and Watts \(1996\)](#). In particular, CL can be structured to incorporate the deep learning approach suggested by [Marton and Saljo \(1976\)](#).

[Johnston et al. \(2001\)](#) evaluated a pair of intermediate-level macroeconomic sections designed with a recitation component. One section had a CL recitation with a deep learning component and the other was a traditional recitation. Recitation participation was required at the cost of one hour of lecture time per week for a 13-week course. The CL recitation students were broken into groups of four students and required to work with each other through various problems: while in the control recitation sections, they simply discussed the same problems with tutor-lead discussion. Their main question was whether deep learning activities designed in their CL recitation sections were superior to their traditional recitation sections. Their finding supports the deep learning process. However, their findings were limited to an increase in student appreciation for the tutors who lead

---

\*All authors are affiliated with the Department of Economics, Indiana University, 105 Wylie Hall, 100 S Woodlawn, Bloomington, IN 47405. The e-mail address for Huynh, Jacho-Chávez, and Self (corresponding author) are: [kphuynh@indiana.edu](mailto:kphuynh@indiana.edu), [djachoch@indiana.edu](mailto:djachoch@indiana.edu), and [jksself@indiana.edu](mailto:jksself@indiana.edu), respectively. The authors are indebted to William E. Becker Jr. and Ed Vytlačil for their comments and suggestions. We also thank Jeffrey Racine for providing us with the computer code to estimate treatment effects. Finally, we thank the CL leaders and students who signed the consent form, without which we could not have done this study. We also appreciate the assistance of Danielle Gunkel and staff at the IUB Human Subjects Office, Donna Rinckel, Erika Knudsen and staff at Indiana University Office of the Registrar, along with Harriet Kenny and the administrative staff in the Department of Economics at Indiana University for assisting us with the data collection. All errors in the paper are our own.

the section.

In general they found mixed improvement in student learning practices and little to no improvement in examination performance or student reported interest in the subject of economics. Considering that course exams are the standard measure of performance, their findings were far from conclusive since they found that CL sessions had no impact on student performance. However, their study was limited due to survey instruments and econometric techniques used. In addition, it is unclear if the recitation, regardless if it is a CL or traditional section, is more productive towards student learning rather than just using class time for lecture only with no recitation time each week.

[Yamarik \(2007\)](#) addresses these criticisms by incorporating a deep learning design into one of his intermediate microeconomics courses by forming CL groups within his class. He then required those groups to be maintained throughout the semester and to conduct problem-solving work through those groups during his lectures. His other section was taught as a traditional lecture base pedagogy. He conducted this experiment over two semesters, Spring 2002 and Fall 2004. By controlling for some student observables, he found that students in the two sections that offered a deep learning design performed better on exams than did students in his other two sections. However, [Yamarik \(2007\)](#) was far from conclusive and the methodology employed had limited ability to adequately control for selection found in the typical student body. Further insights are limited because of heterogeneity in student ability measures and learning abilities found in the broader introductory student body. These measures are generally correlated but the direction of causation is ambiguous.

This study will address these criticisms by controlling for selection and asking whether there is a noticeable performance difference between CL participating students and non-CL participating students. We conducted a quasi-randomized experiment on two large introductory microeconomics class sections at Indiana University's Bloomington-Indiana campus in the Fall 2009 semester. Both sections were taught by the same instructor. We addressed the above questions with non-CL participating students acting as a control group and CL-participating students acting as a treatment group. Self-selection by students into the treatment group and the non-randomization of the two groups is addressed through the estimation method. We also included the Test of Understanding of College Economics 4 (TUCE) as an externally designed performance measure, see [Walstad and Rebeck \(2008\)](#) for a discussion.

## **I. Experiment Design and Data Collection**

The experiment was designed around a 16-week semester with the 16th week reserved for finals. Students were advised through their course syllabus and verbally in class lecture that they can elect to participate in 12 weekly CL sessions. The CL sessions were conducted outside normal lecture hours and there was no loss of normal lecture time. Incentives in the form of course points,

which reduced final examination weight, were introduced to encourage student participation in CL. Students choosing to participate in CL would earn as much as six percent of their possible course points, with each of the 12 sessions counting as 0.5 percent. Those not electing to participate in CL did not receive a reduction in the weight of their final exam and they earned no CL points. Participation in the CL sessions was voluntary, each missed CL session resulted in the return of the 0.5 percent course points to their final exam. Students were allowed to miss up to four CL sessions and still receive some CL credit and reduction of their final exam weight. However, once a student missed five sessions, he or she became ineligible for any course credit for CL and their final exam returned to its original course weight. On the first day of class, the structure of incentives and CL sessions were discussed in-class and they were clearly spelled out in the course syllabus.

The CL session structure incorporated the earlier discussed four-student grouping practice and had two basic activities. Except for the sessions immediately following a semester exam, each CL session would start in their four-student teams working through an assigned small group problem which was followed by an interactive discussion of their CL homework. CL homework was posted on the class webpage in advance of the CL session and students were expected to work through the assignment and bring their suggest answers to their CL session. This allowed interactive exchange to the students and encouraged more comprehensive learning. The CL sessions after the first and second exams had the students form their teams and an interactive discussion of the exam questions was conducted.

At the end of the second lecture students were given an opportunity to sign up for CL. Students only knew the time and room number of each CL session prior to their decision to opt-in to CL. Students were asked to state their top three CL session choices and any academic scheduling conflicts. CL leader assignments were not made public until after student CL session assignments were completed. Each student was randomly placed into a CL session based on the randomization of their top three choices with consideration of the total number of requests for a particular CL session. For those students who could not be assigned a CL session in their top three choices, they were randomly placed in one of the remaining open CL sessions. These procedures were implemented to maximize randomization given the ethical and university policy constraints.

The TUCE was added as an external performance measure and provided opportunity to solicit student characteristics via survey questions. The preTUCE was administered during the second lecture along with a short survey requesting background characteristics. Our study primarily addresses students who complete the course but an important point raised in [Becker and Powers \(2001\)](#) is that a significant number of students do not take both the preTUCE and postTUCE, which resulted in large student course attrition rates. To mitigate attrition, incentives were provided to ensure participation in both the preTUCE and postTUCE. Incentives were similar in structure to CL participation incentives. The difference is that the preTUCE reduction is less than the postTUCE and only applies to the second and third exam. The postTUCE affects the final exam and

the postTUCE was administered during the final course lecture of the semester. An important distinction relative to CL incentives was that participation was mandatory and non-compliance resulted in reduction of three and five percent of a student's possible course grade for failure to take the preTUCE and postTUCE, respectively. However, there was no penalty for completion of either TUCE and answering questions incorrectly. The semester exams were administered on Saturday mornings of the fourth, the ninth, and the 14th week of the semester. University policy allows students to withdraw from the course without penalty or special administrative permission up until the Wednesday prior the second exam (this date is commonly referred to as the drop date). After the drop date, students could only withdraw with written permission from their respective Dean's office.

This experiment's design generated a multitude of performance measures and also collected a rich set of observable characteristics of the students. However, due to the limited scope of this paper, only the necessary performance and observable characteristics are utilized and discussed. This study primarily focuses on three student performance measures. The first measure is based on students' raw final grade comprising the average percentage score received on their four exams with the final exam counting as 1.5 the weight of one of the three semester exams. The raw final grade directly measures individual student understanding of course objectives. The second performance measure is the final course grade as measured by the grading policies established in the syllabus and includes all incentives. The incentives are CL participation (discussed earlier), attendance (students received 0.25 points of an extra credit up to 24 lectures attended), online weekly quizzes with ten allowed attempts, and the earlier discussed TUCE incentive. The third performance measure is the percentage change in the students preTUCE and postTUCE scores.

Student characteristics data consist of gender, age, Indiana University Math placement scores, major, and location of high school matriculation. Student data used is limited to students who signed a human subject consent form and completed the course. This study has been reviewed and approved under the human subject procedures set forth by the Indiana University Bloomington Institutional Review Board (IRB) study number 0905000371. Due to limited scope and space considerations, descriptive statistics are focused only on the raw final grade performance measure, in Table 1.

In our sample about 65 percent of the students were male with an average of 18.78 of which 43 percent were 18 years and younger, 40 percent were 19, and 17 percent were 20 and above. About 48 percent of the students hailed from the state of Indiana, 18 percent from Illinois, 25 percent from the rest of the United States, and nine percent from the rest of the world. Most students, about 48 percent, declared major is University Division (traditionally reserved for undeclared majors and freshman), while the 35 percent were Kelley Business School students, 10 percent were declared College of Arts and Science majors, and seven percent came from other majors. The Indiana University Math placement exams are math skill surveys of students used to place students

Table 1: Descriptive Statistics - Fall 2009

	<i>n</i>	Mean	S.D.	<b>Gender</b>	<i>n</i>	<b>CL Attendance</b>	<i>n</i>
<b>Grade (out of 100)</b>	401	70.70	11.97	<i>Male</i>	261	<i>None</i>	55
<i>CL</i>	259	72.37	11.25	<i>Female</i>	140	< 8	87
<i>NO CL</i>	142	67.65	12.66			≥ 8	259
<b>Math Score (out of 30)</b>	401	17.60	5.00	<b>Declared</b>		<b>High School</b>	
<i>CL</i>	259	17.72	4.98	<b>Major</b>	<i>n</i>	<b>Location</b>	<i>n</i>
<i>NO CL</i>	142	17.37	5.04	<i>UDIV</i>	192	<i>IN</i>	194
<b>Age (in years)</b>	401	18.78	0.98	<i>BUS</i>	140	<i>IL</i>	73
<i>CL</i>	259	18.77	0.98	<i>COAS</i>	42	<i>USA</i>	98
<i>NO CL</i>	142	18.81	0.97	<i>OTHER</i>	27	<i>OTHER</i>	36

Note: UDIV, BUS, COAS, OTHER denotes University Division, Business School, College of Arts of Science, and other faculty, respectively. IN, IL, USA, ROW denotes the location of matriculating high school in Indiana, Illinois, rest of the USA, and rest of the world, respectively.

in the appropriate level introductory math courses. This exam has the benefit of being conducted outside of the course setting and are consistently administered across all Indiana University students who register for undergraduate math courses. The Math placement exam addresses [Ballard and Johnson \(2004\)](#) findings of a positive correlation between a student's score on a test of basic mathematical concepts and introductory economics performance. Finally, about 64.6 percent of the students who signed the waiver participated in CL while 35.4 percent did not. The CL student group had a higher mean raw final grade than non-CL students, 72.4 versus 67.8 percent.

## II. Empirical Methodology and Results

To account for self-selection we use econometric techniques taken from the programme evaluation literature. Our notation follows the potential outcome framework. Let  $t_i$  be the binary treatment indicator for each individual  $i$  in our sample, where  $t_i = 1$  if the student attended at least eight CL sessions and  $t_i = 0$  otherwise. The potential outcomes are denoted by  $y_i(1)$  and  $y_i(0)$ , where  $y_i(1)$  represents the final performance measure when  $t_i = 1$  and  $y_i(0)$  is otherwise. Therefore, the student-level treatment effect is simply  $y_i(1) - y_i(0)$ . However, given the way we defined the treatment, student  $i$  either does or does not receive treatment, and thus only one of these two performance measures is observed. There is a missing data problem, i.e. both performance measures are not observed. In our experiment, we are interested in the population average treatment effect (ATE):  $\tau_{ATE} = E[y_i(1) - y_i(0)]$ , and the average treatment effect on the treated (ATT):  $\tau_{ATT} = E[y_i(1) - y_i(0)|t_i = 1]$ . The ATE measures the expected effect of CL participation on

a randomly drawn student from the class population. Similarly, the ATT is the average effect of active CL participation on those treated student’s final performance measures.

Since attendance of CL participation is voluntary, student  $i$ ’s decision to participate is a function of his or her characteristics,  $x_i$ , such as math score, age category (18 and below, 19, and 20 and above), gender, major, and the location of high school matriculation. Rather than assume students participate in at least eight CL sessions randomly, we assume that their attendance is *unconfounded*, i.e. once students’ characteristics,  $x_i$ , are accounted for, the potential grades  $y_i(1)$  and  $y_i(0)$  are independent of whether or not the student participated in the require minimum number of CL sessions. This assumption is also known as *selection on observables*, because in our framework it rules out the possibility of a student’s decision to participate in at least eight CL sessions to depend on unobserved characteristics that also affect his or her final performance.

In addition to unconfoundedness, we also need individuals with the same characteristics  $x_i$  to be observed in both treated and untreated groups. This *overlap* assumption requires  $0 < p(x) \equiv \Pr(t_i = 1|x_i = x) < 1$ , where  $p(x)$  is the *propensity score*. Finally, we mitigate the possibility that a student’s participation affects the final mark of other students by the randomization of CL assignment previously described.

These assumptions allow for various possible estimation procedures (see [Imbens and Wooldridge \(2009\)](#) for discussion). In this paper, we use [Hirano et al.’s \(2003\)](#) estimators of the ATE and ATT, which are defined below:

$$\widehat{\tau}_{ATE} = \frac{1}{n} \sum_{i=1}^n \frac{[t_i - \widehat{p}(x_i)]y_i}{\widehat{p}(x_i)[1 - \widehat{p}(x_i)]} \quad (2.1)$$

$$\widehat{\tau}_{ATT} = \frac{1}{n_1} \sum_{i=1}^n \frac{[t_i - \widehat{p}(x_i)]y_i}{1 - \widehat{p}(x_i)}, \quad (2.2)$$

where  $y_i \equiv t_i y_i(1) + (1 - t_i) y_i(0)$ , and  $n_1 = 259$  (the total number of treated students). These estimators have been shown to be the most efficient in the sense that their variances are the smallest that can be obtained for estimators of ATE and ATT which rely on unconfoundedness, overlap and random sampling. Implementation of the ATE and ATT requires choosing specific estimators of the propensity score,  $\widehat{p}(x)$ , such as: probit (Probit), logit (Logit), and the complementary log-log (Cloglog) model. These models are implemented using the R code provided by Jeffrey Racine of which the details are described in [Li et al. \(2008\)](#). The results are summarized in [Table 2](#).

The ATE results for all specifications are positive in the range of 4.60 to 4.86 percent, implying that a student participating in CL would see a grade increase of roughly an increment in their grade i.e. a B student goes to B+. The 90 percent confidence interval is constructed using 9999 bootstrap replications and verifies that these ATE are statistically significant. The ATT effect is slightly larger than the ATE for all specifications. The Cloglog model yields a statistically significant ATT at 90 percent level while Logit and Probit do not. An intuitive reason for this result is that the Cloglog link function is asymmetric while Probit and Logit are not. The asymmetry in the Cloglog



Table 2: Result of CL participation on Raw Final Grade

	Probit	Logit	Cloglog
ATE	4.71	4.60	4.86
90% CI	(0.90,7.84)	(0.55,7.85)	(1.05,8.39)
ATT	4.90	4.75	5.19
90% CI	(-0.01,8.70)	(-0.41,8.65)	(0.15,9.46)

Note: The raw final grade is out of 100 percent with 259 students participating in CL and 142 who did not. The average treatment effect (ATE) and average treated on the treated (ATT) are estimated using Probit, Logit, and Cloglog propensity scores. The confidence intervals are constructed using 9999 bootstraps.

functional form prevents it from approaching zero and one, that in turn makes the calculation of (2.1) and (2.2) more stable in each bootstrap replication. Nevertheless, the proportion of the Probit and Logit confidence interval that lies within the negative region is quite small and the estimates are computed without trimming the propensity scores. A graph of the estimated propensity scores for each model is available upon request.

In general, this empirical exercise supports the hypothesis that participation in CL session leads to a positive outcome for students. For robustness, we also calculated the treatment effects for two other performance measures: one, incentivized final grade and two, the change in the TUCE scores. In both cases, there is a statistically positive effect of around 8.64 percent for incentivized final grade and 19 percent increase in the TUCE score. For students, the most salient measure is the incentivized final score as it is the final grade they will receive for the course. The TUCE measures are added as an external check on the performance results. Further details on the incentivized final grade and change in TUCE scores are available from the authors upon request.

### III. Conclusions

In this paper, we construct a quasi-random experiment and use programme evaluation methods to quantify student learning outcomes. We show that ethical constraints that restrict experimental designs and student selection need not restrict rigorous evaluation of student achievement. In this current version we limit the scope to analyze the average effects on performance of CL participation student as compared to the non-CL participating student. Controlling for student selection on observables, we find a strong positive connection between CL participation and student performance. These results provide quantifiable support for continued use of this pedagogy in designing introductory microeconomic courses.

The experimental design has provided an opportunity to collect a rich data set on student performance. We have collected the first wave of data and plan to augment it with subsequent waves

over the next couple of years. This richer data set will allow the use of more powerful estimation techniques than has routinely been available in past studies. Detailed questions on the longitudinal nature of the effects of various pedagogical components, such as CL, on student performance can be extensively explored. Future work needs to address if the effects such as those found in this study consistently occur across student cohorts and time. A more extensive data set and a longitudinal approach will help to address this question and such important aspects as identifying the characteristics of students who most benefit from participation in CL.

Looking further into the dynamics of the benefits of CL, our understanding of the learning process would benefit from considering the potential differences in particular CL sessions that are intrinsic to the CL leader, and what they are. In particular, are there particular characteristics of CL leaders that lead to higher student learning and ultimately higher student performance? Focusing on the benefits to students, we would welcome a better understanding of the perceived benefits of CL from the students' perspective. It would be interesting to see if the perceived benefits correspond to the performance benefits and if these perceptions change over time. Along these lines another important question needs to be addressed: does CL participation lead to more persistence by students to complete the course and lead to any additional incentives to complete their degrees?

Clearly the list of unanswered questions is extensive and the benefit to our understanding of the learning process can be enriched by future work that follows a design similar to that put forth in this study. Administrators are always asking for quantifiable results, which are generally elusive. This study presents an avenue to make those results less elusive, while accounting for students' self selection, heterogeneous characteristics, and working within various institutional constraints.

## References

- Ballard, Charles L. and Marianne F. Johnson**, "Basic Math Skills and Performance in an Introductory Economics Class," *Journal of Economic Education*, 2004, 35 (1), 3–23.
- Becker, William E and Michael Watts**, "Chalk and Talk: A National Survey on Teaching Undergraduate Economics," *American Economic Review*, May 1996, 86 (2), 448–53.
- Becker, William Edward and John R. Powers**, "Student performance, attrition, and class size given missing student data," *Economics of Education Review*, 2001, 20 (4), 377–388.
- Hirano, Keisuke, Guido W. Imbens, and Geert Ridder**, "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," *Econometrica*, 2003, 71 (4), 1161–1189.
- Imbens, Guido W. and Jeffrey M. Wooldridge**, "Recent Developments in the Econometrics of Program Evaluation," *Journal of Economic Literature*, March 2009, 47 (1), 5–86.



**Johnston, Carol G., Jenny N. Lye, Ian M. McDonald, and Richard H. James,** “An Evaluation of Collaborative Problem Solving for Learning Economics,” *Journal of Economic Education*, 2001, 31 (1), 13–29.

**Li, Qi, Jeffrey S. Racine, and Jeffrey Wooldridge,** “Estimating Average Treatment Effects with Continuous and Discrete Covariates: The Case of Swan-Ganz Catherization,” *American Economic Review*, 2008, 98 (2), 357–362.

**Marton, F. and R. Saljo,** “On Qualitative Differences in Learning - 1: Outcome and process,” *British Journal of Educational Psychology*, 1976, 46, 4–11.

**Walstad, William B. and Ken Rebeck,** “The Test of Understanding of College Economics,” *American Economic Review*, May 2008, 98 (2), 547–51.

**Yamarik, Steven,** “Does Cooperative Learning Improve Student Learning Outcomes?,” *Journal of Economic Education*, 2007, 38 (3), 259–277.